Hydrobiologia (2006) 566:365–378

M.T. Furse, D. Hering, K. Brabec, A. Buffagni, L. Sandin & P.F.M. Verdonschot (eds), The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods
DOI 10.1007/s10750-006-0083-6

Comparison of macroinvertebrate sampling methods in Europe

Nikolai Friberg[1,*], Leonard Sandin[2], Mike T. Furse[3], Søren E. Larsen[1], Ralph T. Clarke[3] & Peter Haase[4]

[1]*Department of Freshwater Ecology, National Environmental Research Institute, Vejlsøvej 25, DK-8600 Silkeborg, Denmark*
[2]*Department of Environmental Assessment, Swedish University of Agricultural Sciences, P.O. Box 7050, SE-750 07 Uppsala, Sweden*
[3]*Centre for Ecology and Hydrology, CEH Dorset, Winfrith Technology Centre, Winfrith Newburgh, DT2 8ZD, Dorchester Dorset, UK*
[4]*Research Institute and Natural History Museum Senckenberg, Lochmuehle 2, 63599, Biebergemuend, Germany*
*(*Author for correspondence: E-mail: nfr@dmu.du)*

Key words: macroinvertebrates, metrics, bioassessment, Water Framework Directive, STAR

Abstract

The aim of this study was to describe in detail the national macroinvertebrate sampling methods used and to compare them with a common standard, the STAR-AQEM sampling method. Information on national methods and field data were collected from 11 countries (Austria, Czech Republic, Denmark, France, Germany, Greece, Italy, Latvia, Portugal, Sweden, and UK). The sampling included 22 stream types situated in 11 different Ecoregions. Within each country samples were taken in spring and one additional season (summer or autumn) using both the national method and the STAR-AQEM method. A single anthropogenic stressor was also defined for each stream type sampled within the project, with the three main stressor types being organic pollution (including eutrophication), toxic pollution and habitat degradation. In addition, not impacted reference sites were sampled in each country. A common set of metrics was calculated and compared between the methods. The majority of national methods employed had many features in common. Most of the 12 metrics analysed using the values derived from the STAR-AQEM method and the various national methods correlated significantly, and positively to each other. There was no clear pattern with respect to the differences between metric results obtained using STAR-AQEM and national methods. For some metrics, number of EPT-taxa and families, the value obtained was higher when using the majority of national methods when compared to the STAR-AQEM method. Variability in metric results between methods could not be explained from differences in sampling effort. Sorting in the field and sub-sampling appeared to affect e.g., number of taxa found negatively. The results of the present study supports that inter-calibration in Europe can be undertaken using samples collected with the existing national methods.

Introduction

Macroinvertebrates are the most frequently used organism group in biomonitoring of streams and rivers worldwide (e.g., Metcalfe-Smith, 1996). Currently more than 50 different approaches for biomonitoring using macroinvertebrates exist (De Pauw & Vanhooren, 1983; Metcalfe-Smith, 1996) and most countries in Europe have national and/or regional monitoring programmes that use macroinvertebrates (Birk & Hering, 2003). In most cases each country has developed individual sampling methodology and assessment systems. Among these are RIVPACS (Wright et al., 2000) in the UK, IBGN (AFNOR, 1982) in France and BBI in Belgium (De Pauw & Vanhooren, 1983). As

the majority of methods have evolved from the same ancestors, the Saprobic index and the Trent index (Metcalfe-Smith, 1996), and use a hand net for sampling in accordance with CEN Standard EN 27 828 they should have many features in common. However, until now there has been no direct inter-comparison of the performance of various national sampling methods at a European scale. Johnson et al. (2001) found in an inter-country comparison that national sampling methods in the Nordic countries yielded very similar results.

The EU Water Framework Directive (Directive 2000/60/EC – Establishing a Framework for Community Action in the Field of Water Policy) defines a framework for assessing water bodies including streams and rivers. One of the indicator groups to be used in WFD monitoring of stream and rivers are macroinvertebrates. Intercalibration of the various methods used is essential in providing a consistent picture of the ecological quality within the EU. If existing methods cannot be intercalibrated within an acceptable range of precision, a common standard on sampling methodology should be developed. This is one of the key questions addressed by the STAR project (www.eu-star.at) and the main focus of the present study which was based on data collected from 11 countries (Austria, Czech Republic, Denmark, France, Germany, Greece, Italy, Latvia, Portugal, Sweden, and UK). The sampling included 22 stream types, where five were defined as being of the STAR project type 'Core stream type 1' (mid altitude, 200–500 m.a.s.l., and with a 'small' catchment area 10–100 km^2), seven were of the STAR project type 'Core stream type 2' (lowland, <200 m.a.s.l., and 'medium' catchment areas 100–1000 km^2), whereas ten other stream types were defined as STAR project type 'Additional stream type' (having a different characterisation). These stream types are situated in 11 Ecoregions according to Illies definition (Illies, 1978; as used in the Water Framework Directive), these were regions 3, 4, 6, 7, 8, 9, 10, 14, 15, 16 and 18.

The aim of this study was to describe in detail the national sampling methods used and to compare them against a common standard, the STAR-AQEM sampling method. We hypothesise that sampling effort and subsequent sampling treatment will have a clear impact on the final assessment result. Our additional aim was therefore to elucidate what components of the various methods affected their overall performance.

Material and methods

Sampling strategy

Within each country one or several STAR project stream types were sampled (see Introduction). Macroinvertebrate samples were taken in two different sampling seasons (all partners sampled in spring and one additional season [summer or autumn]). A single anthropogenic stressor was also defined for each stream type sampled within the project, with the two main stressor types being organic pollution (including eutrophication) and habitat degradation (Furse et al., 2006). For each stream type in each country a pre-defined number of sites were selected to cover all ecological classes from high to bad quality (poor quality when the stressor was habitat degradation).

For all sites investigated it were not always possible to apply both the national and the STAR-AQEM sampling method. However, in the comparison of sampling methods data were only included in the analysis where both sampling methods for macroinvertebrates were used at the same site in the same stream and in the same season. The number of samples used for these comparisons therefore differed between types, seasons, and methods used. The analysis undertaken in the present paper combines season and stream types.

Taxonomic adjustment

Each country has adjusted all of its own taxonomical data, so that there are no biases within each country's dataset caused by differences in taxonomic resolution used (e.g., between sampling seasons, where during some seasons it might be more difficult to identify certain taxa because they are in early instars). The taxonomic adjustments were made using common rules within the project.

Comparison of methods

When samples were obtained using a hand net, the area sampled cannot be completely fixed. However,

as the sampling effort should be similar as long as the sampling protocol is followed, the number of individuals obtained should be directly comparable among samples. In addition, the area sampled can be roughly estimated from the area disturbed in front of net multiplied with net width. With respect to the RIVPACS and the PERLA method it is assumed that sampling distance is 1 m per 20 seconds and the area was calculated by multiplying total sampling length (= sampling time/20 s) with net width. Using this approximation the area sampled can be compared among methods. To further enable an inter-method comparison, a 'normalised sampling effort' (NSE) was calculated for each method using sampled area and mesh size. The STAR-AQEM was used as base line and was set to give a NSE value of 1. Consequently, NSE was calculated using the following formula:

$$\mathrm{NSE} = (\mathrm{samplearea}/1.25\mathrm{m}^2)/(\mathrm{meshsize}/0.5\mathrm{mm})$$

as the STAR-AQEM method samples an approximate area of 1.25 m^2 with a 0.5 mm mesh hand net. The NSE is dimensionless. If the method included a pick sample it was not used in the estimation of NSE.

To further allow an inter-comparison of methods used, a handling-processing score was calculated divided into a field and laboratory component. The score is subjective and based on giving the value 1 to each of the handling-processing steps which are considered by the authors to be positive for overall assessment quality (0 if negative), i.e. a high score indicates a high quality method (8 is maximum). In the field, the score 1 is given if field sorting is not undertaken, if no species are removed (1) and if no excess material is removed (1). In the laboratory, the score 1 is given if no live sorting is undertaken (1), if no subsampling is undertaken (1), if sorting is done using magnification (1), if all individuals are enumerated (1) and if identification is done to the species level (1). A sampling method was judged as enumerating all individuals either if it was an actual total count of all individuals in the sample or putting them into abundance classes. In the latter case the actual number will often be based on estimation. Identification to the species level means that all taxa are identified to the best attainable level and that the subsequent index calculation, to which the sampling method was developed, is at least partly based on species information. Project partners supplied all information on the national sampling methods for each country to ensure the most updated information. More details on the various methods can be found in Deliverable 8 of the STAR project, which is published on the STAR homepage (www.eu-star.at).

Metrics used

A group of metrics was selected which was generally applicable and covers various types of stress (e.g., Metcalfe-Smith, 1996; AQEM manual, 2002; Birk & Hering, 2003). The metrics vary in intrinsic properties as to which features of the macroinvertebrate community they respond to, i.e. structural (incl. sensitivity) or functional properties (Table 1). Metric values were calculated from species data using the various national methods and the STAR-AQEM method. This allows for a direct comparison of the national method with the STAR-AQEM method for each country individually.

Statistical analysis

The 12 metrics were calculated from samples obtained using the various national methodologies and the STAR-AQEM method. Only main samples were used (as opposed to replicate samples, where a second sample was taken in some streams to estimate sample variability) so that each site was represented by one sample per season. The national method and the STAR-AQEM method was tested using pair-wise comparisons for each country individually. This was accomplished by performing a Students *t*-test, or a non-parametric Sign test (Sokal & Rohlf, 1995) if the differences in metric values between the STAR-AQEM and national method for a given site and season were not normally distributed. Furthermore the correlation between the STAR-AQEM and national method was investigated by Spearman's rank correlation. For metrics with high correlations, the functional relationship between the STAR-AQEM and national method was investigated and estimated. For a number of selected metrics we plotted their dependence on NSE and the handling-processing score and tested for significant

Table 1. Common metrics used for the comparison of national methods and the STAR-AQEM method

Metric	Type
Saprobic Index (Zelinka & Marvan, 1961)	Structural (sensitivity)
Abundance	Structural – total number of individuals
ASPT (Armitage et al., 1983)	Structural (sensitivity)
Shannon–Wiener index (Shannon & Weaver, 1949)	Structural (diversity)
EPT-taxa	Structural (sensitivity) – total number of taxa belonging to Ephemeroptera, Plecoptera and Trichoptera
No. of taxa	Structural (diversity) – total number of taxa
No. of families	Structural (diversity) – total number of families
Oligochaeta [%]	Structural (insensitivity) – percentage of Oligochaeta in the sample
RETI (Schweder, 1992)	Functional
% Grazers	Functional – percentage of the individuals belonging to functional feeding group grazers in the sample
% Gatherers	Functional – percentage of the individuals belonging to functional feeding group gatherers in the sample
% Shredders	Functional – percentage of the individuals belonging to functional feeding group shredders in the sample

differences using one-way ANOVA followed by a *t*-test (pair-wise comparisons). Box and whisker plots were used for plotting NSE and the handling-processing score versus selected metrics.

Results

Comparison of sampling methods

The majority of sampling methods employed by the different countries have many features in common (Table 2). The majority of methods involve an *a priori* assessment of habitats at the sampling site, exceptions being the RIVPACS method and the DSFI method. In RIVPACS, habitats are sampled in proportion to their occurrence, which is subjectively assessed by the surveyor while sampling. DSFI uses a fixed sampling grid that should cover most habitats without introducing a sampling bias due to variability in how surveyors assess the number of habitats present. All methods, except the Swedish method, use a multi-habitat sampling approach. In contrast, it is the only method, which takes replicate samples to assess inter-sample variability. Most methods use standard hand nets with a width of 25 cm and mesh bag with a 500 μm mesh size in accordance with the CEN standard EN 27 828. The samples are therefore semi-quantitative. A Surber sampler can be used when employing the STAR-AQEM method, while it is obligatory when using the French IBGN protocol with the exception of sampling in lentic areas. Mesh sizes used vary between 475 and 1000 μm. Three of the methods (RIVPACS, DSFI and PERLA) include a pick sample of attached macroinvertebrates.

The smallest area sampled is 0.4 m^2 (IBGN) and the largest is 2.25 m^2 (RIVPACS and PERLA). NSEs ranged from 0.32 (IBGN) to 1.8 (PERLA). Three methods used field sorting of the whole sample (IBE, PERLA and the Latvian method), four collected some species for further identification in the field (STAR-AQEM, IBE, PERLA and the Latvian method) and excess material was removed using most methods. Only when using DSFI, IBGN and PMP is removal of excess material in the field not allowed.

Field sorting is only standard when applying the Italian IBE protocol and the Latvian method (Table 2). In addition, if samples are brought back to the laboratory only the IBE method has live sorting as standard. When using RIVPACS, Portuguese PMP and the Latvian sampling method live sorting is optional, but dead sorting is recommended. All other methods rely on the sorting of dead material. Only the STAR-AQEM method allows sub-sampling of the entire sample. With regard to sorting under magnification, enumeration of all individuals collected and identification to the best attainable taxonomic level, the methods investigated are highly variable. Enumeration of all individuals and identification to the best attainable level increase the biological information in the sample and hence potentially the quality of the assessment. The handling-processing score ranges between 1 (IBE) and 7 (the Swedish method) with most methods obtaining scores of either 4 or 5.

Correlation between STAR-AQEM and National methods

The majority of the 12 metrics analysed using values derived from the STAR-AQEM method and the various national methods correlated significantly and positively to each other (Table 3). Only a few correlations were negative. However, despite being significant a substantial number of correlations had coefficients below 0.7. Overall, number of EPT-taxa was the metric that was most highly correlated when compared among countries. Also the RETI index was highly correlated in most countries. The metric with the overall weakest correlation in an inter-country comparison was abundance. Especially four countries exhibited strong correlations between their national method and the STAR-AQEM method. These were the Czech Republic, Germany, Sweden and the UK. In contrast, especially Italy, but also Denmark and Portugal, had many weak correlations, although some lack of significance can be explained from the low number of sites in these countries. Strong correlations do not necessarily mean that methods will provide identical results. However, they show that results from the different method can be compared.

Table 2. Comparison of the methods used (Y = yes, N = no, V = variable). Calculation of NSE and handling-processing score are explained in the text

		STAR-AQEM	RIVPACS	Nordic methods		The French (IBGN) sampling method	The Italian (IBE) sampling method	The Czech (PERLA) sampling method	The Portugeuse (PMP) sampling method	The Latvian sampling method
				The Swedish method	The Danish (DSFI) method					
Strategy	*A priori* habitat assessment	Y	N	Y	N	Y	Y	Y	Y	Y
	Multi-habitat/number of habitats	Y/V	Y/V	N/1	Y/V	Y/8	Y/V	Y/V	Y/V	Y/V
	No. of samples/replicates	20/none	1/none	5/5	12/none	8/none	1/none	1/none	10/none	20/none
Effort	Sampling device	Hand net or Surber	Hand net	Hand net	Hand net	Surber	Hand net	Hand net	Hand net	Hand net
	Width of sampling device (m)	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.25	0.205
	Mesh size (μm)	500	1000	500	500	500	475	500	500	1000
	Kicking technique (area sample^{-1}, time used, distance sampled)	Fixed area (0.0625 m^2)	Time (3 min)	Time/distance (1 m 1 min^{-1})	Fixed area (0.1 m^2)	Fixed area (0.05 m^2)	Variable	Time (3 min)	Fixed area 1 m sampled (0.205 m^2)	Fixed area 1 m sampled (0.205 m^2)
	Pick sampling (effective sampling time)	N	Y (1 min)	N	Y (5 min)	N	N	N	N	Y
	Area covered (m^2)	1.25	2.25	1.25	1.20	0.4	≈ 0.9[d]	2.25	2.05	4.1
	Sampling effort (NSE)	1	0.9	1	0.96	0.32	0.77	1.8	1.64	1.64

Field	Field sorting	N	N	N	N	N	Y	N	N	Y
	Some species collected from sample in the field	Y	N[a]	N	N	N	Y	Y	N	Y
	Excess material removed	Y	Y	Y	N	N	Y	Y	N	Y
Labora-	Live sorting	N	N[b]	N	N	N	Y	N	N[b]	N[b]
tory	Sub-sampling	Y	N	N	N	N	N	N	N[c]	N
	Use of magnification sorting	N	N	Y	N	Y	N	N	N	N
	Enumeration of all individuals	Y	Y	Y	N	N	N	Y	Y	N
	Identification to species level	Y	N	Y	N	N	N	Y	N	N
	Handling/processing score	4	5	7	5	6	1	5	6	2

[a]Except rare species which are released into the stream or river again, in the case where it is possible to identify them properly in the field and no biomass data is necessary.
[b]Dead sorting is recommended, but live sorting is optional.
[c]Only the fine fraction (>0.5 mm and <1 mm) can be subsampled.
[d]As the number of kick samples varies, the area is set using expert judgement (CNR-IRSA).

Table 3. Correlation matrix between the STAR-AQEM method and the respective national methods

	Abundance	Number of taxa	Saprobic	ASPT	Shannon – Wiener	Grazers (%)	Shredders (%)	Gatherers (%)	RETI	Oligochaeta (%)	EPT – taxa	Number of families
Austria	0.34	0.80***	0.62**	0.63**	0.84***	0.87***	0.80***	0.71**	0.85***	0.69**	0.74***	0.73***
	0.66**	0.71***	0.84***	0.51*	0.70**	0.71***	0.79***	0.72***	0.76***	0.29	0.77***	0.65**
Czech Rebublic	0.63**	0.86***	0.96***	0.89***	0.73***	0.71***	0.93***	0.85***	0.82***	0.80***	0.94***	0.87***
	0.39	0.90***	0.83***	0.90***	0.60**	0.65**	0.94***	0.72***	0.89***	0.87***	0.94***	0.87***
Denmark	0.67*	0.71*	0.82**	0.91***	0.14	0.55	0.28	0.72*	0.70*	0.49	0.72*	0.85**
	0.17	0.85**	0.68*	0.95***	0.76**	0.73*	0.61*	0.87***	0.87***	0.54	0.95***	0.90***
France	0.76*	0.70*	Not possible	0.83**	0.83**	0.87**	0.85**	0.86**	0.83**	0.59	0.85**	0.44
	0.13	0.51	Not possible	0.78**	0.82**	0.81**	0.80**	0.55	0.84**	0.73*	0.83**	0.57
Germany	0.66***	0.51*	0.93***	0.78***	0.68***	0.92***	0.83***	0.76***	0.70***	0.45*	0.86***	0.43*
	0.77***	0.66**	0.89***	0.74***	0.82***	0.87***	0.82***	0.60**	0.72***	0.19	0.65**	0.53*
Greece	0.05	0.93***	Not possible	0.61	0.78*	0.61	0.59	0.77*	0.66*	0.76*	0.79*	0.92***
	0.13	0.60***	Not possible	0.70***	0.79***	0.84***	0.34*	0.67***	0.47*	−0.18	0.64***	0.58***
Italy	0.10	0.49	0.52	0.39	0.56	0.18	0.50	0.35	0.38	No data	0.58	0.35
	−0.26	0.37	0.55	0.73*	0.22	0.63*	0.52	0.67*	0.37	No data	0.65*	0.33
Latvia	0.30	0.62**	0.64**	0.20	0.40	0.49*	0.65**	0.55*	0.73***	0.28	0.46*	0.63**
	0.49*	0.66**	0.79***	0.73***	0.38	0.37	0.85***	0.59**	0.63**	0.58**	0.25	0.69***
Portugal	0.37	−0.18	0.88**	0.75*	0.59	0.30	0.88**	0.88**	0.09	0.67*	0.45	-0.09
	0.54	0.82*	0.84**	0.63	0.72*	0.55	0.92***	0.66	0.76*	0.55	0.56	0.86**
Sweden	0.50*	0.80***	0.85***	0.84***	0.85***	0.82***	0.72***	0.70***	0.82***	0.59**	0.83***	0.74***
	0.67***	0.64***	0.93***	0.91***	0.51*	0.85***	0.68***	0.78***	0.81***	0.53**	0.83***	0.69***
UK	0.82***	0.77***	0.80***	0.88***	0.77***	0.89***	0.83***	0.77***	0.77***	0.62**	0.95***	0.88***
	0.60**	0.90***	0.93***	0.92***	0.66***	0.58**	0.85***	0.64**	0.64**	0.39	0.95***	0.89***

In the top panel for each country are correlations calculated on spring samples and in the lower panel correlations are calculated using summer/autumn samples. Significant correlations are denoted: *$p < 0.05$; **$p < 0.005$: ***$p < 0.0005$. Note that the number of samples varies among countries and between seasons. Therefore, similar correlation coefficients might not have the same p-value.

Table 4. Significant differences between the AQEM-STAR method (S) and the respective national methods (N)

	Abundance	Number of taxa	Saprobic index	ASPT	Shannon–Wiener	%Grazers	%Shredders	%Gatherers	RETI	%Oligochaeta	EPT – taxa	Number of families
Austria					S > N*	S > N*	S < N**				S < N*	S < N*
Czech	S < N*	S < N*	S > N*		S > N**			S > N*		S > N***	S < N***	S < N*
Denmark	S > N***	S < N***									S < N***	S < N***
France	S < N***		S < N***		S > N***						S < N**	S < N***
Germany	S > N*										S < N**	S < N*
Greece		S > N***	N/A									S > N***
Italy	S > N***	S > N***					S < N*			N/A	S > N**	S > N***
Latvia	S > N***	S > N*			S < N***			S > N***		S > N***		S > N***
Portugal		S < N**									S < N*	S < N*
Sweden		S < N***				S < N*	S < N*				S < N*	S < N***
UK	S > N***		S > N*		S > N**	S < N*	S < N***	S > N***	S < N***	S > N***		S < N*

Analyses are based on all sites and seasons within each country.*$p < 0.05$; **$p < 0.005$; ***$p < 0.0005$.

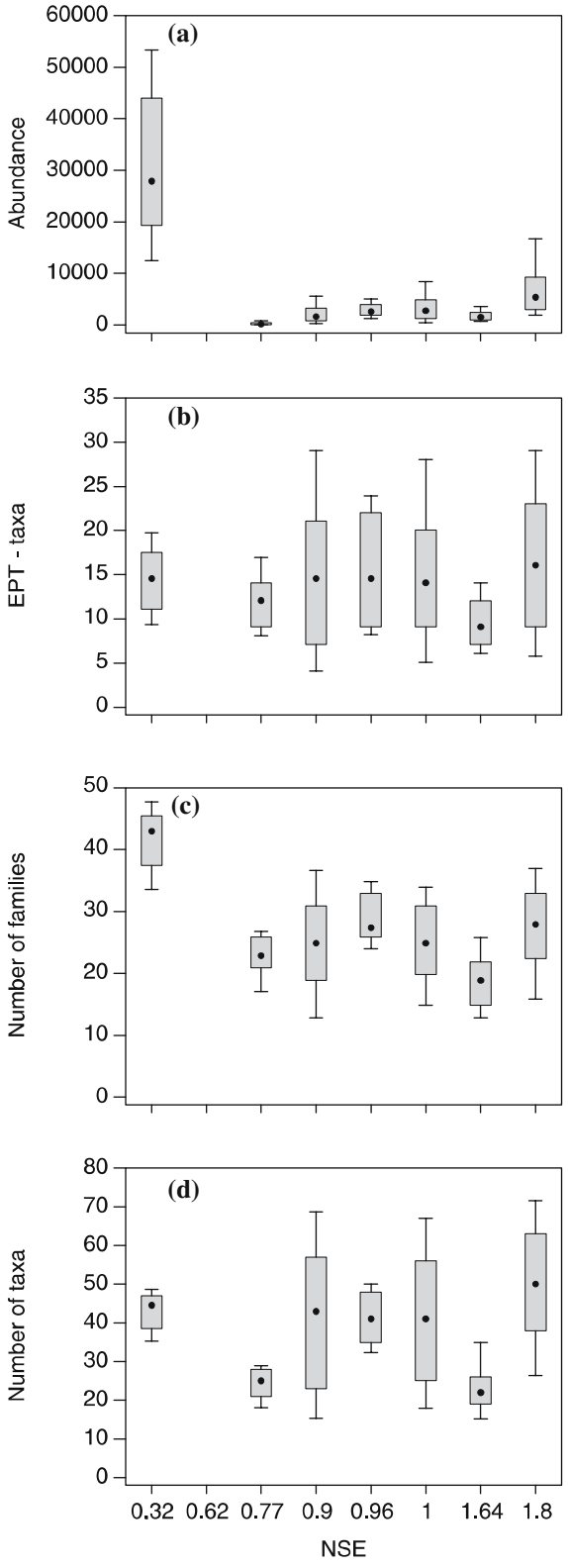

Figure 1. Relationship between normalised sampling effort (NSE) and abundance (a), EPT-taxa (b), number of families (c) and number of taxa (d). Median values (circle), 75th and 25th percentile (top and bottom edge of box, respectively) and 90th and 10th percentile (top and bottom of error bars, respectively) are shown.

Comparison of AQEM-STAR and national methods

No overall clear pattern emerged with respect to the differences between metric results obtained using STAR-AQEM and national methods (Table 4). Within countries, there was, in most cases, not a consistent pattern when comparing metrics: some metrics would score higher when calculated using data obtained by the national method while other would score lower than the STAR-AQEM method.

In most cases (64% of the countries) the various national methods yielded significantly higher EPT-taxa values than the STAR-AQEM method. A similar pattern was evident with respect to number of families. In 73% of the countries significantly more families were found using the national method. In contrast, the STAR-AQEM method yielded significantly more EPT-taxa and families in 9 and 27% of the countries, respectively.

The STAR-AQEM method yielded in general higher values (e.g. more taxa) than the national methods in Italy and Latvia when the methods were significantly different whereas the opposite was the case in Sweden and Portugal where the national method consistently yielded higher metric values than the STAR-AQEM method. In Denmark and Germany, significantly more individuals were found when employing the STAR-AQEM method whereas the opposite was true with respect to number of EPT-taxa and families.

Several countries used the RIVPACS method as their national method (Austria, Germany, Greece and UK; Table 2). In addition, the Czech PERLA system is very closely related to the RIVPACS method (Table 2). Overall, there were no clearly consistent results among these countries.

Inter-country comparison of metric performance

There was no relationship between abundance of macroinvertebrates in samples and NSE (Fig. 1a).

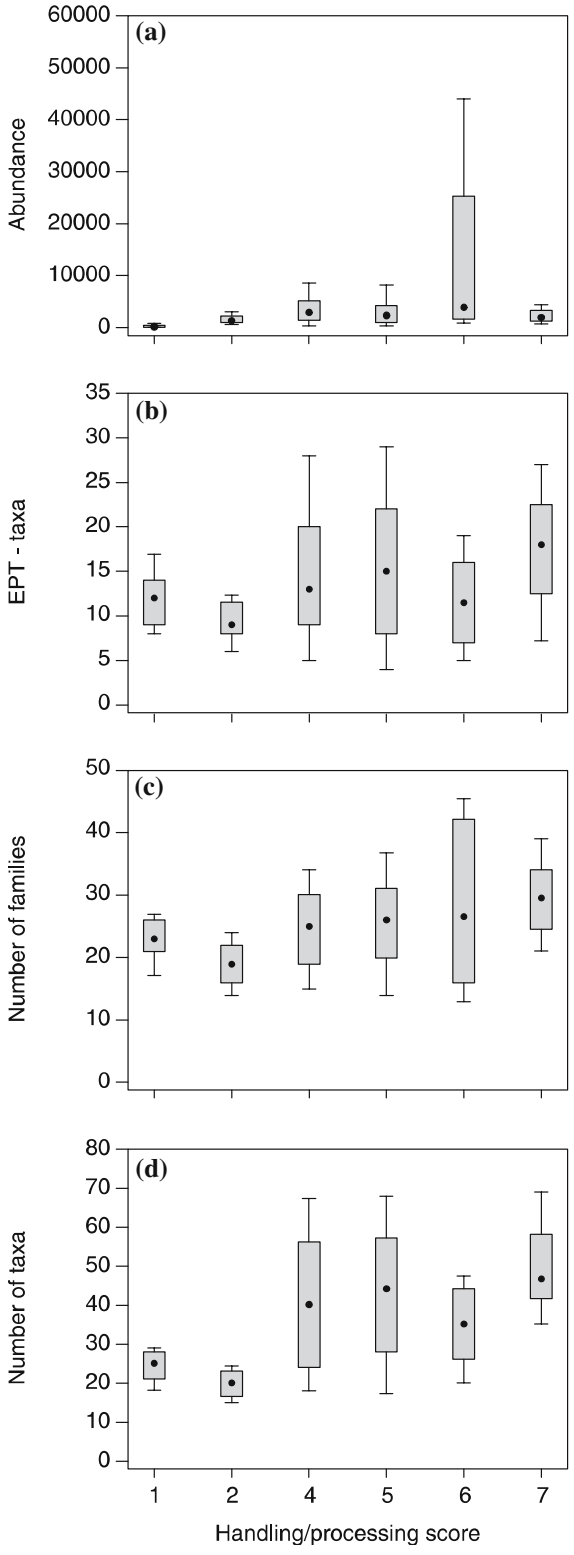

Figure 2. Relationship between handling/processing score and abundance (a), EPT-taxa (b), number of families (c) and number of taxa (d). Median values (circle), 75th and 25th percentile (top and bottom edge of box, respectively) and 90th and 10th percentile (top and bottom of error bars, respectively) are shown.

The French methods IBGN had a significant higher number of individuals than all methods and at the same time the lowest NSE ($p < 0.0001$). If the IBGN method is omitted from the data set there is a tendency for an increase in number of individuals caught with increasing NSE. There was no clear relationship between the number of EPT-taxa and NSE (Fig. 1b). The IBGN method caught a similar number of EPT-taxa as the other methods despite the low NSE. The number of families found was not related to NSE (Fig. 1c). As with abundance, the method with the smallest sampling area and NSE caught significantly the largest number of families (the IBGN method, $p < 0.0001$). The number of taxa was, as for the other metrics tested, not related to NSE (Fig. 1d). There was a high degree of variability, which appears to be method specific and cannot be explained from single variables as NSE.

Abundance was lower in samples with a handling/processing score of 1 (the IBE method) whereas abundance varied independently of the score in the range 4–7 (Fig. 2a), except for the IBGN method catching much more individuals than the other methods with a handling score of 6. There was a tendency that the number of EPT-taxa found increased with increasing handling/processing score, indicating that these taxa are lost during sample treatment (Fig. 2b). There was no effect of the handling/processing score on the number of families found (Fig. 2c) whereas there was significantly fewer taxa found when scores were 1 and 2 compared with scores 4–7 (Fig. 2d, $p < 0.0001$).

Discussion

The national methods compared in the present study had many features in common. All methods, except the French IBGN method, used a hand net and in most cases the mesh size was 500 μm in

accordance with the CEN standard EN 27 828. It is therefore not surprising that the various methods yielded comparable results. That different sampling methods will provide almost identical results have previously been demonstrated in the Nordic countries (Johnson et al., 2001). They found that sampling methods from four countries (Denmark, Finland, Norway and Sweden) yielded very similar results when sampling was done in one perturbed and one unperturbed stream in south-central Sweden. Another study also shows similar results between RIVPACS and STAR-AQEM (Haase et al. 2004a). The STAR-AQEM method appeared to collect more individuals and taxa than the national methods in Italy and Latvia. This could reflect the very low handling-processing score obtained for both countries compared with the STAR-AQEM method as well as the other national methods. With respect to Latvia, a further explanation could be that a number of taxa are not considered in the national method, and consequently they will not appear in the taxa list. In Sweden and Portugal, the national method yielded consistently more taxa, EPT-taxa and families than the STAR-AQEM method. This could relate to the use of subsampling in the STAR-AQEM methodology, which might reduce the number of taxa found. In the case of Sweden, the higher number of taxa (all and EPT) and families might reflect that the sampling effort is concentrated in riffles which are the most species rich in stream ecosystems (e.g., Brown & Brussock, 2001). In Denmark and Germany, significantly more individuals were found when employing the STAR-AQEM method whereas the opposite was true with respect to number of EPT-taxa and families. Again, this might reflect that taxa are lost when subsampling the large STAR-AQEM sample.

Several countries used the RIVPACS method as their national method (Austria, Germany, Greece and UK; Table 2). In addition, the Czech PERLA system is very closely related to the RIVPACS method. Overall, there were no clearly consistent results among these countries. As the differences cannot be attributed to protocol itself, they might reflect the way samples were taken in the individual countries. It has previously been shown that sampling potentially is a major source of variation when employing the RIVPACS techniques (Clark, 2000; Dines & Murray-Bligh, 2000). In the STAR project, a workshop was undertaken prior to the start of the sampling programme in which the various methods, including the RIVPACS methodology, were demonstrated in order to reduce sampling variability among countries. This might not have been sufficient in reducing the variability as our results indicate that differences among countries in how sampling is undertaken are as important as the intrinsic differences in the methods employed.

Handling in the field and processing of samples in the laboratory will affect the quality of the assessment result. Field sorting, collection of some species from the sample in the field and removal of excess material can all potentially reduce sample quality by the loss of species (Haase et al., 2004b). Field handling is extremely dependent on the surveyors' abilities and is affected by weather conditions, time pressure etc. However, removal of fragile or endangered species can be necessary in certain cases and any negative impacts on sample quality should be reduced through training of the surveyors (e.g., Dines & Murray-Bligh, 2000). Obligatory live sorting is likely to affect quality negatively as it introduces a time constraint on the sorting procedure. Even though this is not obvious from the NSE value of the STAR-AQEM method, it collects large amounts of inorganic material, organic debris and plants, which makes subsampling necessary. Sub-sampling can potentially reduce the number of species found and hence affect sample quality negatively and increase sampling variance (Vinson & Hawkins, 1996). Sorting under magnification increases the likelihood of finding all species present in the sample, even the smaller specimens.

In conclusion, the STAR-AQEM method appears to collect fewer taxa (all and EPT) and families than the majority of the national methods. The most likely explanation of this finding is that species are lost during the sub-sampling procedure employed by the STAR-AQEM method. However, an additional explanation could be that the STAR-AQEM was developed to take samples habitat proportional, ignoring rare habitats which might contain additional species (AQEM, 2002). The advantage of this approach is that it limits sampling variability by reducing the subjective element introduced by the surveyor and that it is likely to be more sensitive towards hydromor-

phological degradation. Therefore, the lower number of taxa found in the STAR-AQEM samples than in the national methods might to some degree reflect a higher sensitivity to hydromorphological degradation. Two methods, the Italian IBE method and the Latvian method, appear to lose information about the macroinvertebrate community to a degree that might affect the assessment of ecological stream quality. Laboratory processing (IBE and Latvian method) and identification of more species (Latvian method) would probably improve their performance. Despite these differences it is difficult to estimate the effects of different methods on assessment results. Differences in single metrics might be covered by a multi-metric approach. In Germany, for example, the assessment results (multi-metric system with scores between 0 and 1) are highly correlated (Spearman $R = 0.92$) and their differences very small (mean difference −0.01), when comparing STAR-AQEM and RIVPACS (Haase et al., 2004a).

The results of the present study are promising as it clearly indicates that existing national sampling methods can be relatively easily intercalibrated as they are in general based on similar principles. It consequently supports that intercalibration among European countries at presently is undertaken by calculating a common set of metrics on species lists collected by the individual countries using their national methodology (Buffagni et al. 2005). However, it is important to keep in mind that many aspects of sampling methodology, such as sensitivity to different stressors and stability in time and space, was not covered by the present analysis. Many of the national methods were developed to detect organic pollution, focusing on the collection of indicator species that might occur in rare habitats. However, pressures on stream ecosystems change in time and consequently also the stressors acting on the biota. As examples, hydromorphology and introduction of exotic species are increasingly important and will affect assessment results using existing methods/systems in an inconsistent manner (e.g., Olsen & Friberg, 1999; Gabriels et al., 2005). Consequently, most methods should be improved in several aspects and the revision and improvement of methods should be an ongoing process.

Acknowledgement

STAR was partially funded by the European Commission, 5th Framework Program, Energy, Environment and Sustainable Development, Key Action Water, Contract no. EVK1-CT-2001-00089. The authors thank the whole STAR team for making this paper possible, especially the colleagues who collected the field samples and sorted and identified the taxa.

References

AFNOR, 1982. Essais des euax. Détermination de l'indice biologique global normalisé (IBGN) Association Francaise de Normalisation NF T 90–350 France.

AQEM, 2002. Manual for the application of the AQEM system, version 1.0. www.aqem.de.

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research 17: 333–347.

Birk S. & D. Hering, 2003. Waterview Web-Database: a comprehensive review of European assessment methods for rivers. FBA News, 20 (winter 2002): 4.

Brown, A. V. & P. P. Brussock, 2001. Comparisons of benthic invertebrates between riffles and pools. Hydrobiologia 220: 99–108.

Buffagni, A., S. Erba, S. Birk, M. Cazzola, C. K. Feld T. Ofenböck, J. Murray-Bligh, M. T. Furse, R. Clarke, D. Hering, H. Soszka & W. van der Bund, 2005. Torwards European Inter-Calibration for the Water Framework Directive: Procedures and Examples for the Different River Types from the E.C. Project STAR. 11th STAR Deliverable. STAR contract no: EVK1-CT 2001–00089. Quaderni Instituto di Ricerca sulle Acque 123, IRSA, Rome, Italy 460 pp.

Clarke, R., 2000. Uncertainty in estimates of biological quality based on RIVPACS. In Wright, J. F., D. W. Sutcliffe, M. T. Furse (eds), Assessing the Biological Quality of Freshwaters: RIVPACS and Similar Techniques. Freshwater Biological Association, Ambleside, UK, 39–54.

De Pauw, N. & G. Vanhooren, 1983. Method for biological quality assessment of watercourses in Belgium. Hydrobiologia 100: 153–168.

Dines, R. A. & J. A. D. Murray-Bligh, 2000. Quality assurance and RIVPACS. In Wright, J. F., D. W. Sutcliffe & M. T. Furse (eds), Assessing the Biological Quality of Freshwaters: RIVPACS and Similar Techniques. Freshwater Biological Association, Ambleside, UK, 71–78.

Gabriels, W., P. L. M. Goethals & N. De Pauw, 2005. Implications of taxonomic modifications and alien species on biological water quality assessment as exemplified by the Belgian Biotic Index method. Hydrobiologia 542: 137–150.

Haase, P., S. Lohse, S. Pauls, K. Schindehütte, A. Sundermann, P. Rolauffs & D. Hering, 2004a. Assessing streams in Ger-

many with benthic invertebrates: development of a practical standardised protocol for macroinvertebrate sampling and sorting. Limnologica 34: 349–365.

Haase, P., S. Pauls, A. Sundermann & A. Zenker, 2004b. Testing different sorting techniques in macroinvertebrate samples from running waters. Limnologica 34: 366–378.

Johnson, R. K., K. Aagaard, K. J. Aanes, N. Friberg, G. M. Gislason, H. Lax & L. Sandin, 2001. Macroinvertebrates. In Skriver, J. (ed.), Biological Monitoring in Nordic Rivers and Lakes. TemaNord 2001:513, Nordic Council of Ministers, Copenhagen, Denmark: 43–52.

Metcalfe-Smith, J. L., 1996. Biological water-quality assessment of rivers: use of macroinvertebrate communities. In Petts, G. & P. Calow (eds), River Restoration. Blackwell Science, Oxford, UK, 17–59.

Olsen, H. -M. & N. Friberg, 1999. Biological stream assessment in Denmark: The importance of physical factors. In Friberg, N. & J. D. Carl (eds), Proceedings from the Second Meeting of the Nordic Benthological Society, Silkeborg November 1997. National Environmental Research Institute, Denmark.

Schweder, H., 1992. Neue indices für die Bewertung des ökologischen Zustandes von Fliessgewässern, abgeleitet aus der Makroinvertebraten-Ernährungstypologie. Limnologie Aktuell 3: 353–377.

Shannon, C. E. & W. Weaver, 1949. The Mathematical Theory of Communication. The University of Illinois Press, Urbana, IL, U.S.

Sokal, R. R. & J. R. Rohlf, 1995. Biometry. Freeman and Company, New York, U.S.

Vinson, M. R. & C. P. Hawkins, 1996. Effects of sampling area and subsampling procedure on comparisons of taxa richness among streams. Journal of the North American Benthological Society 15: 392–399.

Wright, J. F., D. W. Sutcliffe & M. T. Furse (eds), 2000. Assessing the Biological Quality of Freshwaters: RIVPACS and Similar Techniques. Freshwater Biological Association, Ambleside, UK.

Zelinka, M. & P. Marvan, 1961. Zur Präzisierung der biologischen Klassifikation der Reinheit fließender Gewässer. Archiv Fur Hydrobiologie 57: 389–407.